FROM NOISY NEURAL TIME SERIES TO STRUCTURED LANGUAGE: A FOUNDATION MODEL FOR IMAGINED SPEECH DECODING FROM EEG SIGNALS

Sparsh Rastogi[1], **Kanav Dhanda**[1], **Akshat Bakshi**[1], **Tanmay Kumar**[2], **Jatin Bedi**[1]
[1] Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, India
[2] University of New South Wales, Australia
{srastogi_be22, kdhanda_be23, abakshi_be23, jatin.bedi}@thapar.edu
z5696466@ad.unsw.edu.au

ABSTRACT

Communicative brain–computer interfaces (BCIs) offer a promising pathway for restoring communication in patients affected by conditions such as amyotrophic lateral sclerosis (ALS). Among these paradigms, imagined speech decoding from non-invasive EEG is particularly attractive due to its portability and scalability. However, EEG constitutes a highly noisy neural time series, characterized by low signal-to-noise ratios, substantial inter-subject variability, and susceptibility to artifacts. Moreover, imagined speech arises from purely internal cognitive processes, producing weak and spatially diffused neural activity. Extracting structured semantic information from such signals remains a significant challenge.

To address this challenge, we present NeuroSpeak, a JEPA-based framework for sentence-level imagined speech generation from non-invasive EEG. Our approach combines masked neural signal modeling with vector-quantized latent discretization to learn robust EEG representations, which are aligned with language embeddings using a predictive alignment objective and decoded into natural language via a pretrained sequence model. We train and evaluate our model on the large-scale CHISCO corpus comprising over 20,000 imagined speech sentences under a subject-agnostic evaluation setting. The proposed framework achieves a semantic similarity score of 47.70% relative to ground-truth text, demonstrating generalization beyond subject-specific neural patterns. To the best of our knowledge, this represents the largest and most semantically diverse study of sentence-level imagined speech generation using non-invasive EEG.

1 INTRODUCTION

Human speech serves as a primary channel for conveying semantic information, generated through tightly coordinated neural, articulatory, and respiratory processes. Disruptions to these pathways, as observed in conditions such as amyotrophic lateral sclerosis (ALS) or locked-in syndrome, can severely restrict communication, motivating the development of Brain–Computer Interfaces (BCIs) that decode language directly from neural activity. Recent advances in neural decoding using convolutional, recurrent, and transformer-based architectures have demonstrated impressive results, especially when leveraging invasive modalities such as electrocorticography (ECoG) Angrick et al. (2019); Willett et al. (2023); Metzger et al. (2023). These approaches benefit from high resolution, enabling speech synthesis from cortical recordings. However, their reliance on surgically implanted electrodes limits scalability and broader applicability. Less invasive approaches use stereoelectroencephalography (sEEG) Angrick et al. (2022); Wu et al. (2024); Chen et al. (2025) which reduces surgical burden but still requires implantation. Consequently, there has been a growing interest in non-invasive modalities, like electroencephalography (EEG) and magnetoencephalography (MEG) for imagine speech decoding. MEG-based systems have demonstrated promising results in acoustic reconstruction and speech synthesis Dash et al. (2020); Verkhlyutov et al. (2023); Kwon et al. (2024), but remain constrained by high costs and limited portability. EEG, while significantly more deployable, poses a substantially more challenging modeling problem. It constitutes a high-dimensional,

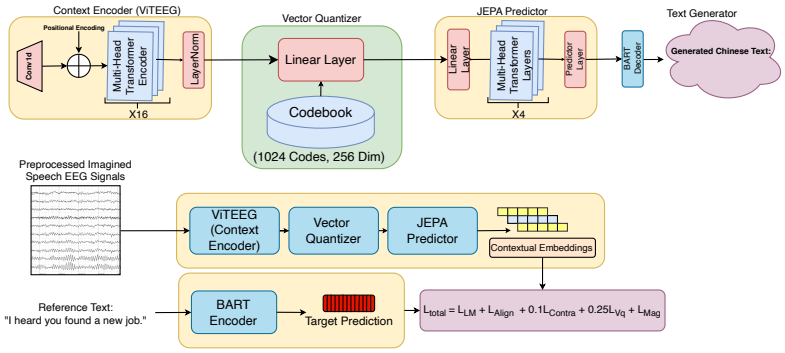

Figure 1: Illustration of the proposed framework.

low signal-to-noise multivariate neural time series marked by pronounced inter-subject variability and susceptibility to physiological and environmental artifacts. Furthermore, imagined speech arises from purely internal cognitive processes, producing neural activity that is weaker and more spatially diffuse than that observed during overt speech or reading Nalborczyk et al. (2023). Recovering structured semantic information from such signals therefore remains a fundamental challenge. Existing approaches to imagined speech decoding from non-invasive EEG largely operate at the word Asghari Bejestani et al. (2022); Xiong et al. (2025) or phoneme level Varshney & Khan (2022), typically framing the task as classification. While effective under constrained vocabularies, these methods struggle to capture the compositional and semantic structure of language and do not naturally extend to sentence-level generation. More recent sentence-level frameworks Rastogi et al. (2025) remain limited by small-scale datasets, restricting the ability of models to learn robust and generalizable linguistic representations.

In this work, we present early results from an ongoing effort to develop a cross-subject generalizable framework for sentence-level imagined speech generation from non-invasive EEG. We formulate imagined speech decoding as a representation learning problem over noisy neural time series. Our approach combines masked signal modeling with vector-quantized latent discretization to extract stable latent representations from raw EEG signals. These neural embeddings are then aligned with language representations using a Joint Embedding Predictive Architecture (JEPA) objective, enabling the recovery of structured semantic information, and subsequently decoded into natural language using a pretrained sequence model. We train and evaluate the framework on the CHISCO corpus, a large-scale Chinese imagined speech dataset comprising over 20,000 sentences and more than 900 minutes of high-density EEG recordings per participant across 39 semantic categories. To assess robustness, we adopt a subject-agnostic evaluation setting that measures generalization to unseen sentences and participants. To the best of our knowledge, this constitutes the largest and most semantically diverse study of sentence-level imagined speech generation using non-invasive EEG. Our key contributions could be summarised as follows:

- We propose a foundational model for sentence-level imagined speech generation leveraging masked signal modeling, vector-quantized latent discretization, and JEPA-based cross-modal alignment to recover structured semantic representations from noisy EEG signals.
- We conduct large-scale training and subject-agnostic evaluation on over 20,000 imagined speech sentences, explicitly assessing cross-sentence and cross-subject generalization and demonstrating robust sentence-level generation beyond prior EEG-based approaches.

2 METHODOLOGY

The architecture of our framework is illustrated in Fig. 1. Our model receives a 125-channel EEG signal $X \in \mathbb{R}^{C \times T}$ as input, where C is the number of channels and T is the number of time steps. The model broadly consists of the following major components.

Feature Extraction: Given raw EEG signals, we first apply a one-dimensional convolutional layer to process the multi-channel time series and extract low-level neural features. These features are augmented with positional encodings and passed to a multi-head Vision Transformer–based encoder, which models global temporal dependencies via self-attention to produce contextualized representations for semantic alignment and text generation. To further strengthen representation learning, we adopt a masked modeling strategy inspired by Rastogi et al. (2025), wherein a large proportion of EEG tokens are randomly masked and reconstructed from the remaining context. This objective encourages the encoder to capture global structure and learn robust latent representations.

Vector Codebook Discretization: The embeddings generated in the previous stage are projected into a lower-dimensional space to suppress noise and emphasize task-relevant neural signals. These compressed representations are used to train a learnable discrete vector codebook; based on empirical analysis, we adopt a codebook of 1024 vectors to balance representational capacity and stability. Each embedding is assigned to its nearest codebook entry via approximate nearest-neighbor search, and the codebook is optimized using a contrastive loss (Eq. 1) to capture shared structure and diversity across samples. This discretization encourages the model to learn shared latent representations for semantically similar neural patterns, facilitating structured and semantically coherent alignment with language. The selected codebook vectors are then projected back to the original embedding dimension for subsequent processing while preserving the learned discrete semantic structure.

$$\mathcal{L}_{\text{Contra}} = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_{\text{pred}}^{(i)}, \mathbf{z}_{\text{teacher}}^{(i)}\right)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}\left(\mathbf{z}_{\text{pred}}^{(i)}, \mathbf{z}_{\text{teacher}}^{(j)}\right)/\tau\right)} \tag{1}$$

JEPA Predictor To align EEG-derived representations with linguistic semantics, we employ a Joint Embedding Predictive Alignment (JEPA) objective within a student–teacher framework. The student, implemented as a multi-head transformer, receives the selected codebook embedding and predicts a target language embedding generated by a frozen teacher model. Training minimizes a similarity-based objective between the predicted and teacher embeddings (Eq. 2), encouraging the model to capture high-level semantic structure rather than low-level neural signal variations. During inference, the predicted EEG embedding conditions a pretrained language decoder for text generation. The embedding is provided as encoder-side contextual input, enabling sentence generation (Eq. 3). This decoupling of semantic alignment from language modeling allows the system to leverage strong linguistic priors while relying solely on EEG-derived representations for semantic conditioning.

We used the Chinese Imagined Speech Corpus (CHISCO), a large-scale non-invasive EEG dataset for sentence-level imagined speech decoding Zhang et al. (2024), comprising over 20,000 sentences from five healthy participants, each contributing more than 900 minutes of high-density EEG recordings across 39 semantic categories. Each trial consists of a reading phase (5.0 s) followed by an imagined speech phase (3.3 s), and we use the imagined speech interval for modeling. The dataset is released with standardized preprocessing that preserves broadband EEG information while attenuating artifacts through re-referencing, filtering, and automated artifact rejection. For modeling, imagined speech segments are extracted into fixed-length windows with consistent channel ordering across subjects and standardized per channel using training-set statistics. To ensure rigorous evaluation (Fig. 2), we adopt a leave-one-subject-out protocol in which each participant is held out once for testing, resulting in five evaluation folds. Within each fold, approximately 20% of sentences shared across the remaining subjects are reserved for validation, enforcing disjointness at both the subject and lexical levels. Final evaluation is conducted on the held-out subject, including an additional probe restricted to unseen sentences to assess cross-subject and lexical generalization.

$$\mathcal{L}_{\text{Align}} = 1 - \frac{\mathbf{z}_{\text{pred}} \cdot \mathbf{z}_{\text{teacher}}}{\|\mathbf{z}_{\text{pred}}\|_2 \, \|\mathbf{z}_{\text{teacher}}\|_2} \tag{2}$$

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \mathbf{z}_{\text{pred}}) \tag{3}$$

3 Results and Discussion

Various backbone architectures, including Vision Transformer (ViT), Swin Transformer Liu et al. (2021), Data-efficient Image Transformers (DeiT) Touvron et al. (2020), and EfficientNet, were evaluated for extracting EEG feature embeddings. Performance was measured using standard language generation metrics, including BLEU, ROUGE-L, METEOR, perplexity (PPL), semantic similarity,

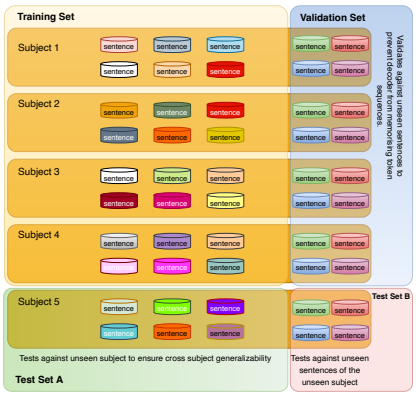

Figure 2: Dataset splits used in our experiments. Orange denotes the training set; blue indicates the validation set with unseen sentences; green corresponds to Test A with unseen subjects; and red represents Test B containing both unseen subjects and unseen sentences.

Table 1: Performance of Various Models under Unified and Separate Vector Quantization Settings

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	PPL	ROUGE-L	METEOR	CERR	SemSim	WER
EfficientNet (Baseline)	0.3159	0.0890	0.0240	0.0018	183.08	0.3472	0.2050	1.3191	0.1920	1.5800
EfficientNet + Unified VQ	0.4054	0.1263	0.0709	0.0406	1.1982	0.4352	0.2449	0.7307	0.3818	1.1301
8-Head ViT + Unified VQ	0.4043	0.1127	0.0578	0.0356	1.1733	0.4469	0.2495	0.7495	0.3854	1.1578
16-Head ViT + Unified VQ	0.4108	0.1163	0.0622	0.0396	1.1740	0.4439	0.2347	0.6440	0.4724	1.0294
DeiT + Separate VQ	0.3860	0.1281	0.0732	0.0436	1.1705	0.4215	0.2517	0.8876	0.3085	1.2944
Swin + Separate VQ	0.3811	0.1170	0.0652	0.0402	1.1717	0.4203	0.2489	0.8908	0.4063	1.2986
16-Head ViT + Separate VQ	0.4201	0.1384	0.0757	0.0416	1.1680	0.4453	0.2544	0.6981	0.2952	1.0994

character error rate (CER), and word error rate (WER). We first evaluated a participant-specific vector quantization (VQ) setting, where separate codebooks were learned per subject to model subject-dependent neural patterns. As shown in Table 1, the 16-head ViT achieved the best performance in this setting, outperforming Swin and DeiT with higher BLEU-1 (0.4201), improved ROUGE-L (0.4453), and lower perplexity (1.1680), indicating stronger lexical alignment and fluency. To enhance robustness and cross-subject generalization, we adopted a unified VQ codebook shared across participants. This setting yielded consistent improvements for ViT-based models. Specifically, the 16-head ViT with unified VQ achieved higher semantic similarity (0.4724 vs. 0.2952), lower CER (0.6440 vs. 0.6981), and lower WER (1.0294 vs. 1.0994) compared to the participant-specific setting, suggesting that shared discretization mitigated subject-specific noise and promoted common semantic representations. Within this unified framework, ViT-based models also consistently outperformed the EfficientNet backbone used in prior work Rastogi et al. (2025), particularly on semantic similarity and error-based metrics, highlighting the advantage of transformer architectures for modeling long-range temporal dependencies in EEG signals. Moreover, the original encoder–decoder architecture from Rastogi et al. (2025) degraded in higher-vocabulary settings due to representational collapse, whereas the proposed JEPA-based framework maintained stable and semantically coherent representations, underscoring the benefit of predictive alignment.

To further assess generalization under inter-subject variability, we conducted leave-one-subject-out evaluation. As shown in Table 2, performance remained stable across all five folds, with BLEU-1 ranging from 0.3810 to 0.4192 and perplexity consistently around 1.17–1.18. Although Subject 3 achieved the highest semantic similarity (0.4513) and lowest WER (1.0203), and Subject 2 presented the most challenging case, performance degradation was limited. Even in the most difficult fold, the model generated coherent sentences, indicating that the learned representations generalized beyond subject-specific neural characteristics. The low variance across folds further confirmed the effectiveness of unified discretization and JEPA-based alignment for cross-subject semantic consistency.

Table 2: Cross-Subject Generalization Performance under Leave-One-Subject-Out Evaluation

Subject	BLEU-1	BLEU-2	BLEU-3	BLEU-4	PPL	ROUGE-L	METEOR	CERR	SemSim	WER
Subject 1 (F)	0.4192	0.1357	0.0741	0.0413	1.1794	0.4447	0.2535	0.7020	0.3291	1.1055
Subject 2 (M)	0.3810	0.1160	0.0652	0.0405	1.1746	0.4202	0.2486	0.8905	0.4053	1.2991
Subject 3 (M)	0.4121	0.1268	0.0726	0.0477	1.1726	0.4469	0.2381	0.6402	0.4513	1.0203
Subject 4 (M)	0.4075	0.1332	0.0754	0.0432	1.1801	0.4428	0.2506	0.7734	0.3541	1.1890
Subject 5 (F)	0.4043	0.1127	0.0578	0.0356	1.1733	0.4469	0.2495	0.7495	0.3854	1.1578

4 CONCLUSION

In this work, we presented NeuroSpeak, a JEPA-based framework that extracts rich semantic representations from noisy time series. Through large-scale evaluation on the CHISCO corpus, we showed that transformer-based encoders combined with unified vector quantization and predictive alignment enable the recovery of semantically meaningful language representations from non-invasive EEG. Under a stringent leave-one-subject-out protocol, the model achieved a semantic similarity score of 47.70%, demonstrating robust cross-subject generalization despite substantial inter-subject variability. These findings suggest that structured linguistic semantics can be recovered from high-dimensional, low signal-to-noise neural time series. As ongoing work, we aim to further strengthen representation robustness, scale evaluations, and extend the framework to multilingual settings. More broadly, this study highlights the potential of JEPA-style modeling for learning transferable semantic representations from noisy time series.

REFERENCES

M Angrick, C Herff, E Mugler, MC Tate, MW Slutzky, DJ Krusienski, and T Schultz. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of Neural Engineering*, 16(3):036019, 2019.

M Angrick, M Ottenhoff, L Diener, D Ivucic, G Ivucic, S Goulis, AJ Colon, L Wagner, DJ Krusienski, PL Kubben, T Schultz, and C Herff. Towards closed-loop speech synthesis from stereotactic eeg: A unit selection approach. In *ICASSP*, pp. 1296–1300, 2022.

MR Asghari Bejestani, Gh R Mohammad Khani, VR Nafisi, and F Darakeh. Eeg-based multiword imagined speech classification for persian words. *BioMed Research International*, 2022(1): 8333084, 2022.

J Chen, X Chen, R Wang, C Le, A Khalilian-Gourtani, E Jensen, P Dugan, W Doyle, O Devinsky, D Friedman, et al. Transformer-based neural speech decoding from surface and depth electrode signals. *Journal of Neural Engineering*, 2025.

Debadatta Dash, P Ferrari, and J Wang. Decoding imagined and spoken phrases from non-invasive neuromagnetic (meg) signals. *Frontiers in Neuroscience*, 14:290, 2020.

J Kwon, D Harwath, D Dash, P Ferrari, and J Wang. Direct speech synthesis from non-invasive, neuromagnetic signals. In *Proc. Interspeech*, pp. 412–416, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. URL https://arxiv.org/abs/2103.14030.

SL Metzger et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, 2023.

Ladislas Nalborczyk, Marieke Longcamp, Mireille Bonnard, Victor Serveau, Laure Spieser, and F.-Xavier Alario. Distinct neural mechanisms support inner speaking and inner hearing. *Cortex*, 169:161–173, 2023. ISSN 0010-9452. doi: https://doi.org/10.1016/j.cortex.2023.09.007. URL https://www.sciencedirect.com/science/article/pii/S0010945223002332.

Sparsh Rastogi, Harsh Dadwal, Khushboo Modi, Jatin Bedi, and Jasmeet Singh. Towards sentence level imagined speech generation from eeg signals. In *Proc. Interspeech*, 2025.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *CoRR*, abs/2012.12877, 2020. URL https://arxiv.org/abs/2012.12877.

Yash V Varshney and Azizuddin Khan. Imagined speech classification using six phonetically distributed words. *Frontiers in Signal Processing*, 2:760643, 2022.

V Verkhlyutov, V Vvedensky, K Gurtovoy, E Burlakov, and O Martynova. Speech recognition from meg data using covariance filters. In *Biologically Inspired Cognitive Architectures Meeting*, pp. 904–911, 2023.

Francis R Willett et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

X Wu, S Wellington, Z Fu, and D Zhang. Speech decoding from stereo-electroencephalography (seeg) signals using advanced deep learning methods. *Journal of Neural Engineering*, 21(3): 036055, 2024.

Wenjing Xiong, Lin Ma, and Haifeng Li. Synthesizing intelligible utterances from eeg of imagined speech. *Frontiers in Neuroscience*, Volume 19 - 2025, 2025. ISSN 1662-453X. doi: 10.3389/fnins.2025.1565848. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2025.1565848.

Zihan Zhang, Xiao Ding, Yu Bao, Yi Zhao, Xia Liang, Bing Qin, and Ting Liu. Chisco: An eeg-based bci dataset for decoding of imagined speech. *Scientific Data*, 11(1265), 2024.